Melodic Transcription of Flamenco Singing from Monophonic and Polyphonic Music Recordings

EMILIA GÓMEZ
JORDI BONADA
JUSTIN SALAMON
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{emilia.gomez, jordi.bonada, justin.salamon}@upf.edu - http://mtg.upf.edu

Abstract

We propose a method for the automatic transcription of flamenco singing from monophonic and polyphonic music recordings. Our transcription system is based on estimating the fundamental frequency (f0) of the singing voice, and follows an iterative strategy for note segmentation and labelling. The generated transcriptions are used in the context of melodic similarity, style classification and pattern detection. In our study, we discuss the difficulties found in transcribing flamenco singing and in evaluating the obtained transcriptions, we analyze the influence of the different steps of the algorithm, and we state the main limitations of our approach and discuss the challenges for future studies.

1. Introduction

1.1 Motivation

Flamenco is a music tradition originating mostly from Andalusia in southern Spain. The origin and evolution of the different flamenco styles (*palos*) and variants have been studied by different disciplines, including ethnomusicology, literature and anthropology. A frequent topic of discussion among flamenco scholars concerns the comparison of different performances and the precise definition of styles and variants. Current Music Information Retrieval (MIR) technologies could provide a different perspective to this debate. First, they might help to set up a standard methodology for flamenco description and comparative analysis, and support the formalization of expert knowledge. Second, they might facilitate the study of large music collections.

Flamenco music germinated and nourished mainly from the singing tradition. Accordingly, the singer's role soon became dominant and fundamental. Often, the singer is accompanied by the flamenco guitar; other flamenco instruments include claps, rhythmic feet and percussion.

This work focuses on melodic description, and is driven by the research hypothesis that each flamenco style is characterized by a certain melodic skeleton or contour, which can be subject to ornamentation and variation (Donnier, 1997; Mora et al., 2010). The aim of this work is therefore to provide a method to extract detailed melodic transcriptions from audio signals, which can then be processed for ornament detection and further simplified to obtain the overall melodic contour. We will deal with a cappella singing (*martinete* and *debla* styles) and singing with guitar accompaniment (*fandango* style).

1.2 Flamenco and its musical transcription

Because of its oral transmission, there are no written scores in flamenco music. Flamenco experts have put much effort into generating manual transcriptions after listening to live performances or field recordings, as a means to catalogue, classify and imitate the most relevant performers and their stylistic traits (Hurtado and Hurtado, 1998); (Hurtado and Hurtado, 2002); (Fernández, 2004); (Hoces, 2011). As pointed out by Toiviainen & Eerola (2006) and Lesaffre et al. (2004) in other contexts, manual analyses provide very accurate and expert information, but they are very time consuming and might be subjective or prone to errors. This is also the case in flamenco, due to two main reasons. First, there is a disagreement on the most adequate transcription methodology; For instance, Donnier (1997) proposed the adaptation of plainchant neumes. Hurtado and Hurtado (1998, 2002), on the contrary, forcefully argue for the use of Western notation. Second, even if we agree on the use of a certain format, there is a degree of subjectivity in the transcription process, given the high degree of ornamentation in flamenco music.

1.3 Automatic transcription of sung melodies

Automatic transcription is one of the main research challenges in the field of sound and music computing. It consists in computing a symbolic musical representation (in terms of Western notation) from a given musical performance (Klapuri, 2006). For monophonic music material, the obtained transcription relates to the melody (Gómez et al., 2003) and in polyphonic music material there is an interest in transcribing the predominant melodic line (Klapuri, 2006). Transcription systems can provide melodic descriptors at different levels. The main melody-related Low-level features are energy, associated with loudness, and fundamental frequency (f0) related to its perceptual correlate, pitch. From now on, we will use the term pitch to refer to f0. In a higher structural level, audio streams are segmented into notes, and their duration and pitch provide a symbolic representation. This representation can be the input to higher-level music analyses, e.g. ornament detection, melodic contour extraction or key or scale analysis. Current systems for automatic transcription are usually composed of three different stages: low-level (frame-based) descriptor extraction, note segmentation and note labelling.

When dealing with monophonic music signals, existing transcription systems provide satisfying results for a great number of musical instruments. Although we find some successful approaches for singing voice (Mulder et al. 2003; Ryynänen, 2006), it is still one of the most complex instruments to transcribe, even in a monophonic context. This is due to several factors, such as the continuous character of the human voice and the variety of pitch ranges and timbre. This results in difficulties in obtaining correct f0 estimations, detecting note transitions and labelling notes in terms of pitch or duration. When dealing with polyphonic music signals, current state-of-the-art algorithms for predominant f0 estimation yield an overall accuracy around 75% according to the 2011 edition of the Music Information Retrieval Evaluation eXchange (MIREX). Moreover, audio onset detection methods yield an average F-measure around 0.78 (MIREX). This F-measure is obtained for a mixed dataset of 85 files, but if we just consider the 5 tested singing voice excerpts, the maximum F-measure is 0.47. In addition, current approaches are oriented towards mainstream popular music. This leads us to the question of how would these algorithms perform for, e.g. traditional music, and more particularly, flamenco singing. Additional challenges in flamenco transcription arise from the quality of existing recordings, the acoustic and expressive particularities of singing, its ornamental and improvisational character and the yet to be formalized musical structures employed (Mora et al., 2010).

2. Selected approach

Figure 1 shows an overall diagram of the proposed system, which is based on the one described in (Janer et al., 2008). It consists of four main steps: low-level feature extraction (fundamental frequency, energy and spectral features), tuning frequency estimation, transcription into short notes, and an iterative process involving note consolidation and refinement of the tuning frequency.

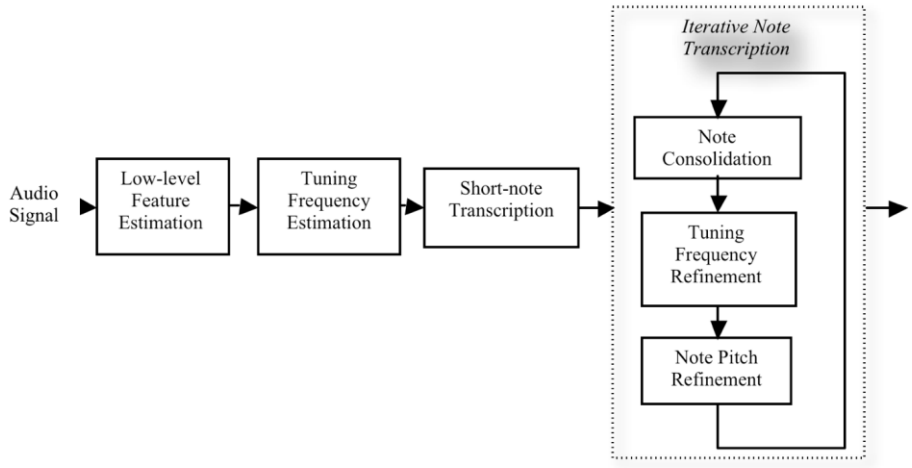

Figure 1: Steps for automatic transcription

2.1 Fundamental frequency estimation from monophonic signals

For a cappella singing, we have evaluated three state-of-the-art approaches for fundamental frequency estimation: 1. Time-domain autocorrelation: we have considered the well-known yin algorithm proposed by de Cheveigné and Kawahara (2002) (*yin*); 2. Frequency-domain harmonic matching: we have implemented an algorithm based on the Two-Way Mismatch algorithm (*twm*) proposed by Maher and Beauchamp (1994), as presented in (Cano, 1998). This algorithm tries to match the spectral peaks (local maxima of the spectrum) to a harmonic series; 3. Frequency-domain autocorrelation: the third method we consider (*sac*: Spectrum Autocorrelation) is based on the computation of amplitude correlation in the frequency domain.

In order to measure the amount of errors caused by wrong f0 estimation in the final performance, we have also introduced a manually edited f0 envelope (*Corrected-f0*). This envelope was obtained by manual edition of the last approach (*sac*), where we manually corrected the most relevant f0 errors, mainly caused by reverberation (end of phrases) and noise (background voices and percussion).

2.2 Predominant fundamental frequency estimation from polyphonic signals

For predominant f0 estimation, we make use of the algorithm by Salamon and Gómez (2012). This algorithm obtained the highest overall accuracy in the most recent MIREX evaluation campaign (Salamon and Gómez, 2011). In the first stage of the algorithm, the audio signal is analyzed and spectral peaks (sinusoids) are extracted. This process is comprised of three main steps: first a time-domain equal loudness filter is applied (Vickers, 2001), which has been shown to attenuate spectral components belonging primarily to non-melody sources (Salamon et al., 2011). Next, the short-time Fourier transform is computed with a 46 ms Hann window, a hop size of 2.9 ms and a x4 zero padding-factor. At each frame the local maxima (peaks) of the spectrum are detected. In the third step, the estimation of the spectral peaks' frequency and amplitude is refined by calculating each peak's instantaneous frequency (IF) using the phase vocoder method (Flanagan and Golden, 1966) and re-estimating its amplitude based on the IF. The detected spectral peaks are subsequently used to compute a representation of pitch salience over time: a *salience function*. The salience function is based on harmonic summation with magnitude weighting, and spans a range of almost five octaves from 55Hz to 1760Hz. Further details are provided in (Salamon et al., 2011).

In the next stage, the peaks of the salience function are grouped over time using heuristics based on auditory streaming cues (Bregman, 1990). This results in a set of pitch contours, out of which the contours belonging to the melody need to be selected. The contours are automatically analyzed and a set of contour characteristics is computed. In the final stage of the system, the contour characteristics and their distributions are used to filter out non-melody contours. The

distribution of contour salience is used to filter out pitch contours at segments of the song where the melody is not present. Given the remaining contours, we compute a rough estimation of the melodic pitch trajectory by averaging at each frame the pitch of all contours present in that frame, and then smoothing the result over time using a sliding mean filter. This mean pitch trajectory is used to minimise octave errors (contours with the correct pitch class but in the wrong octave) and remove pitch outliers (contours representing highly unlikely jumps in the melody). Finally, the melody f0 at each frame is selected out of the remaining pitch contours based on their salience. A full description of the melody extraction algorithm, including a thorough evaluation, is provided in (Salamon and Gómez, 2012).

In addition to computing the melody f0 sequence using the default algorithm parameters, we also computed the sequences adjusting three parameters of the algorithm for each excerpt: the minimum frequency threshold, the maximum frequency threshold and the strictness of the voicing filter (c.f. Salamon and Gómez, 2012 for details about the voicing filter). The results using the adjusted parameters are referred to as *SalamonGomez-adaptedparam*.

2.3 Tuning frequency estimation

As we analyze singing voice performances, the reference frequency used by the singer to tune the piece is unknown. In order to locate the main pitches, we perform an initial estimation of this tuning frequency assuming an equal-tempered scale. We also assume that this reference frequency is constant for the analyzed excerpt. We estimate it by computing the maximum of the histogram of f0 deviations from an equal-tempered scale tuned to 440 Hz. This histogram represents the mapping of the f0 values of all frames into a single semitone interval with a one cent resolution. In our approach, we give more weight to frames where the included f0 is stable by assigning higher weights to frames where the values of the f0 derivative are low. In order to smooth the resulting histogram and improve its robustness to noisy f0 estimations, instead of adding a value to a single bin, we use a bell-shaped window that spans several bins. The maximum of this histogram ($b_{\max}$) determines the tuning frequency deviation in cents from 440 Hz. Therefore, the estimated tuning frequency in Hz becomes $f_{ref} = 440 \cdot 2^{b_{\max}/1200}$.

2.4 Short note transcription

The f0 sequence is then segmented into short notes by using a dynamic programming (DP) algorithm based on finding the segmentation that maximizes a set of probability functions. The estimated segmentation corresponds to the optimal path among all possible paths along a 2-D matrix M (see Figure 2). This matrix has the possible note pitches in cents as rows ($[c_0, c_n]$) and the analysis frame times as columns. Note that the possible note pitches should cover the tessitura of the singer ($[c_{\min}, c_{\max}]$) and include a $-\infty$ value for the unvoiced sections. In this step, note durations are limited to a certain range between $n_{\min}$ and $n_{\max}$ frames. The maximum duration $n_{\max}$ should be long enough so that it covers several periods of a vibrato with a low modulation frequency, e.g. 2.5 Hz, but also short enough as to have a good temporal resolution, for example, a resolution that avoids skipping fast notes with a very short duration.

Possible paths considered by the DP algorithm always start from the first frame, end at the last audio frame, and advance in time so that notes never overlap. A path P is defined by its sequence of m notes, $P = \{N_0, N_1, \ldots, N_{m-1}\}$, where each note N_i begins at a certain frame k_i, has a pitch deviation of c_i in cents relative to the tuning reference, and a duration of n_i frames. The optimal path is defined as the path with maximum likelihood among all possible paths. The likelihood L_P of a certain path is determined as the product of likelihoods of each note (L_{N_i}) times the likelihood of each jump between consecutive notes (L_{N_{i-1},N_i}), that is

$$L_P = L_{N_0} \cdot \prod_{i=1}^{m-1} L_{N_i} \cdot L_{N_{i-1},N_i} .$$

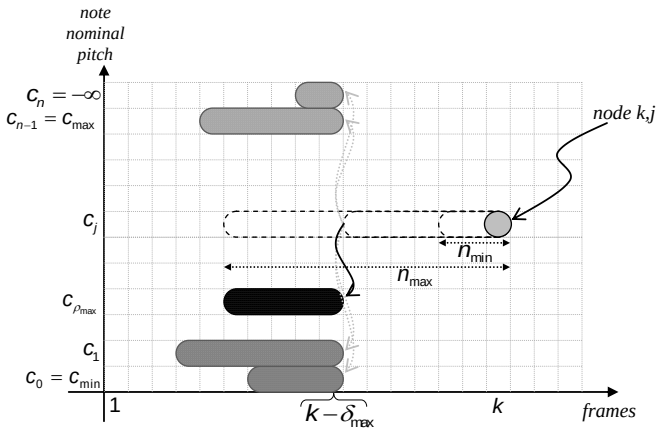

Figure 2: This figure shows the matrix *M* used by the short note segmentation process, and illustrates how the best path for the node with frame k and note j is determined. All possible note durations between n_{mix} and $n_{\max}$ are considered, as well as all possible jumps to previous notes. In this example $\delta_{\max}$ is found to be the most likely note duration and $\rho_{\max}$ the index of the previous note.

In our approach, no particular characteristic is assumed *a priori* for the sung melody; therefore all possible note jumps have the same likelihood $L_{N_{i-1},N_i} = 1,\ \forall i \in [1, c_n - 1]$. On the other hand, the likelihood L_{N_i} of a note N_i is determined as the product of several likelihood functions based on the following criteria: duration (L_{dur}), fundamental frequency (L_{pitch}), existence of voiced and unvoiced frames ($L_{voicing}$), and low-level features related to stability ($L_{stability}$). For a note N_i, its likelihood L_{N_i} is computed as $L_{N_i} = L_{dur} \cdot L_{pitch} \cdot L_{voicing} \cdot L_{stability}$. Duration likelihood L_{dur} is set so that it is small for short and long durations. Pitch likelihood L_{pitch} is defined so that the likelihood is higher the closer the estimated pitch contour values are to the note nominal pitch c_i and vice versa, giving more relevance to frames with lower values for the first derivative of the pitch contour. The voicing likelihood $L_{voicing}$ is defined so that segments with a high percentage of unvoiced frames are unlikely to be a voiced note, while segments with a high percentage of voiced frames are unlikely to be an unvoiced note. Finally, the stability likelihood considers that a voiced note is unlikely to have fast and significant timbre or energy changes in the middle. Note that this is not in contradiction with the typical characteristic of flamenco singing of changing the vowel at ending notes, since those changes are mostly smooth.

2.5 Iterative note consolidation and tuning frequency refinement

In the last step, consecutive notes with the same pitch and a smooth transition are consolidated, the estimated tuning frequency is refined according to the obtained notes, and the note nominal pitch is re-estimated based on the new tuning frequency. This whole process is repeated until there are no more consolidations.

Note consolidation: the notes obtained in the previous step have a limited duration between $n_{\min}$ and $n_{\max}$, although longer notes are likely to have been sung. Therefore, it makes sense to consolidate consecutive voiced notes into longer notes if they have the same pitch. However, significant and fast energy or timbre changes around the note connection boundary may be indicative of phonetic changes unlikely to happen within a note, and thus may indicate that those consecutive notes are different ones. Thus, consecutive notes will be consolidated only if they have the same pitch and the stability measure of their connection falls below a certain threshold.

Tuning frequency refinement: In a previous step, tuning frequency was estimated from the fundamental frequency contour. However, once notes have been segmented, it may be beneficial to use the note segmentation to refine the tuning frequency. For this purpose, we compute a

pitch deviation for each voiced note, and then estimate the new tuning frequency from a one-semitone histogram of weighted note pitch deviations. Weights are determined as a measure of the salience of each note, giving more weight to longer and louder notes.

Figure 3 shows an example of the note transcription of a monophonic music recording. The system transcribes according to an equal-tempered scale, as requested by flamenco experts. That means that, even if the performer is out of tune, we approximate the used scale to a chromatic scale, i.e., mistuning is not transcribed.

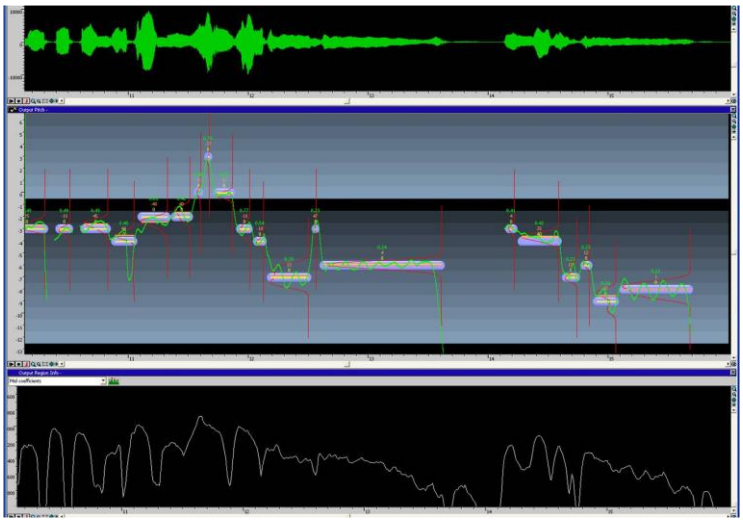

Figure 3: Example of the visualization tool for melodic transcription. Audio waveform (top), estimated f0 and pitch (middle) and energy (bottom).

3. Evaluation strategy

3.1 Music collections

For monophonic music recordings, we gathered a music collection of 72 sung excerpts representative of different a cappella singing styles (*Tonás*). This collection was built in the context of a study on similarity and style classification of flamenco a cappella singing styles. We refer to (Mora et al. 2010) for a comprehensive description of the considered styles and their musical characteristics. All 72 excerpts are monophonic, their average duration is 30 seconds and there is enough variability for a proper evaluation of our methods, including a variety of singers, recording conditions, presence of percussion, clapping, background voices and noise. The files contain a total of 211047 frames and 2803 notes. In addition, we built a small control data set of pop/jazz a cappella singing, consisting of 5 musical phrases by different singers, recorded in good conditions. This control dataset will serve to evaluate the difficulty in transcribing flamenco material and test the algorithms in easier conditions to establish a performance ceiling.

For polyphonic music recordings, we gathered approximately 20 minutes of music, consisting of 20 excerpts of singing voice with guitar accompaniment (*Fandango* style). This collection was also built in the context of the COFLA project[1]. It too contains a variety of male and female singers and recording conditions. The average duration of the analyzed samples is 57 seconds and the files contain a total of 189454 frames and 1680 notes.

3.2 Ground truth gathering

We collected manual annotations from a musician with limited knowledge of flamenco music, so that there was no implicit knowledge applied in the transcription process. In order to gather manual annotations, we provided a user interface for visualizing the waveform and fundamental frequency in cents (in a piano roll), as shown in Figure 3. Since transcribing everything from

[1] http://mtg.upf.edu/research/projects/cofla

scratch is very time consuming, we also provided the output of a baseline transcription method based on manually corrected fundamental frequency values (*corrected f0*). The annotator could listen to the original waveform and the synthesized transcription, while editing the melodic data until he was satisfied with the transcription. The criteria used to differentiate ornaments and pitch glides were discussed with two flamenco experts, so that the annotator could follow a well-defined strategy.

3.3 Evaluation measures

We computed several evaluation measures as done in the (MIREX) *Audio Melody Extraction* task, which are derived from the comparison of frame-based f0 values and pitch values, comparing the algorithm's output against the ground truth reference. To evaluate voicing detection, we compute two measures: voicing recall, i.e. the percentage of voiced frames according to the reference that are declared as voiced by the algorithm; and voicing false alarm, i.e. the percentage of unvoiced frames according to the reference that are declared as voiced by the algorithm.

To evaluate pitch accuracy, we compute the raw pitch accuracy, i.e. the percentage of voiced frames where the pitch estimation is correct, considering a certain tolerance or threshold in cents (*th*). This threshold is needed given the fact that frequency values are quantized to (equal-tempered) pitch values with respect to the estimated tuning frequency. This results in a small mistuning of the estimated fundamental frequency envelopes. We evaluate the raw pitch accuracy using different threshold values.

We also compute two global accuracy measures, defined as follows. Raw chroma accuracy is defined as the percentage of voiced frames where the chroma estimation is correct, considering a certain tolerance or threshold in cents (*th*). This measure allows for octave error in the estimation. Finally, the overall accuracy represents the percentage of frames that have been correctly estimated in terms of pitch (for voiced frames) or correctly detected as unvoiced frames.

We assessed the influence of two main steps of the algorithm on the evaluation results. First, we analyzed the effect of the f0 estimation method by comparing several algorithms and a manually edited f0 envelope, as described above. Second, we considered the influence of note segmentation by comparing our approach with an alternative method (*mami*) proposed by Mulder et al. (2003) for monophonic music material. For this second step, we did not have access to the f0 envelope, but only to the final melodic transcription output.

4. Results

4.1 Melodic transcription from monophonic music recordings

We start by analyzing the f0 and note transcription outputs from a cappella singing (i.e. monophonic). The first thing we note is that there is a small detuning between the outputs of the algorithm when using different monophonic f0 estimation algorithms,. This occurs since the tuning frequency is estimated based on the f0 values. This detuning in turn results in small transpositions between the estimated transcriptions. In addition, the algorithm assumes a constant tuning with respect to 440 Hz. We observe that, for some excerpts, the tuning varies along time, meaning this assumption does not hold. This results in further detuning in the transcriptions. For these reasons, it is important to compare the obtained representations considering this small detuning. We also see that different algorithms produce dissimilar segmentation results in short notes, as illustrated in Figure 4. This is due to the fact that the note consolidation procedure highly depends on input f0 envelope.

The evaluation results for a cappella singing and for different tolerance intervals are provided in Figure 5. When considering a tolerance of 100 cents (1 semitone), the segmentation algorithm proposed in this study yields the best overall accuracy when using the corrected f0 envelope (90.43%), followed by *sac* (81.68%) and *twm* (79.14%). The results for the compared state-of-the-art approach (*mami)* are also very close to our system (79.1%).

The worst results are given by the *yin* algorithm (68.56%). In terms of pitch accuracy, the proposed approach (using either *sac* or *twm* for f0 estimation) outperforms *mami* and *yin*. We believe this is probably due to the fact that both the *sac* and *twm* algorithms have been specially designed for singing voice. We also observe that *mami* has a better voicing false alarm (*vx_false_alm_av*) than our approach. This fact together with the high difference in voicing false alarm between *corrected_f0* and *sac* or *twm* indicates that the system would benefit from an improved voicing detection procedure after f0 estimation.

If we decrease the tolerance to half a semitone (50 cents), the overall accuracy decreases for all the considered approaches (e.g. 76.81% for *corrected f0*, 69.56% for *sac*). The ranking of algorithms is also similar to that obtained using the 100 cents tolerance, although the accuracy of the *mami* approach is closer to *yin*. Finally, for a 25 cents tolerance, the ranking of methods is almost the same but with the difference that *mami* obtains the lowest *overall accuracy*. This might be due to the fact that *mami* does not quantize note pitches to an equal-tempered scale, as done in the ground truth. As expected, the overall accuracy increases with the tolerance for all the considered approaches.

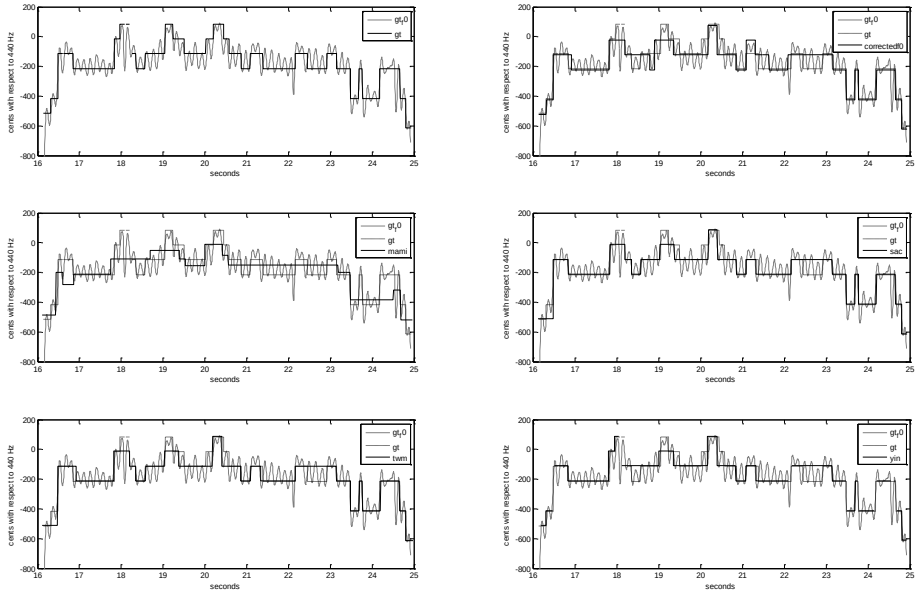

Figure 4: Examples of frame-based note transcription together with the fundamental frequency envelope (*corrected_f0*) in an excerpt for a Debla style (by Naranjito).

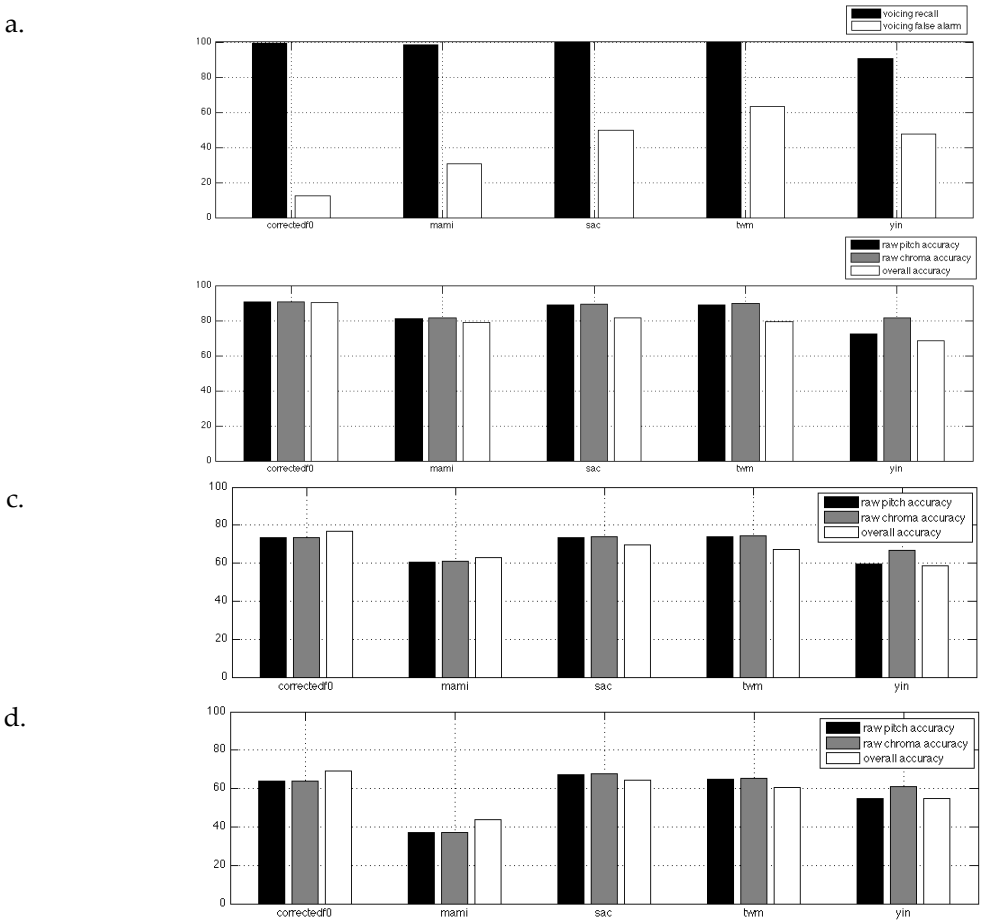

Figure 5: Frame-based accuracy measures: voicing recall and voicing false alarm (a) and accuracy measures with 100 cents (b), 50 cents (c) and 25 cents (d) tolerance.

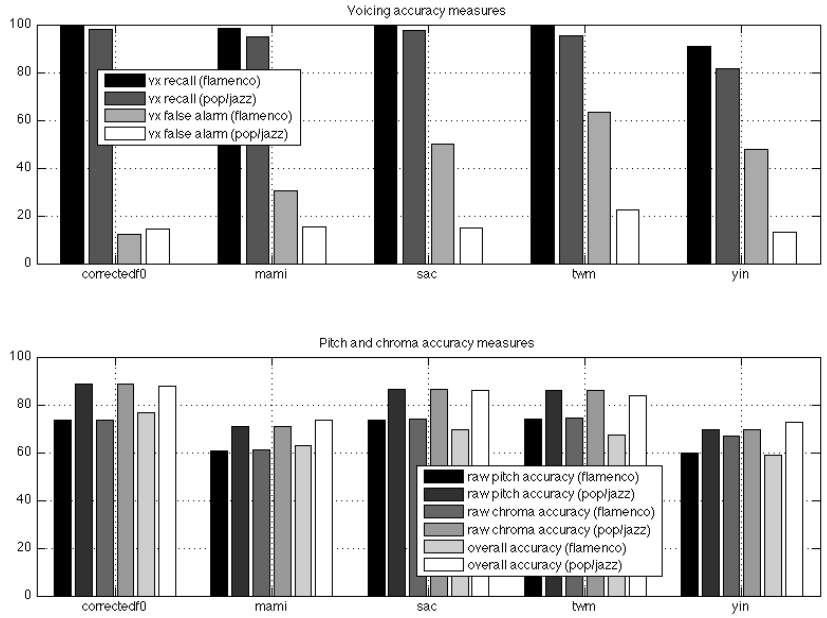

Figure 6: Comparison of frame-based accuracy measures (50 cents tolerance) for flamenco against a control dataset from pop/jazz.

In Figure 6 we provide a comparison of the accuracy for flamenco vs rock/jazz singing. We observe that the highest overall accuracy obtained for pop/jazz singing is 87%. Although the pop/jazz material if very limited for a representative evaluation, this result suggests that the aforementioned methods indeed work better for these singing styles.

4.2 Melodic transcription from polyphonic music recordings

For singing with guitar accompaniment, we observe in Figure 7 that the overall accuracy for our approach is 72.6%, which is around 3% higher than for the monophonic case (*sac* algorithm). This is due to two main reasons. First, as the voice is very predominant with respect to the guitar, the f0 estimation method works very well for this material. Second, as the singer follows the tuning reference of the guitar, there are no tuning errors and the note segmentation results are improved. This overall accuracy is improved to 83.6% if we adapt the f0 estimation parameters for each considered excerpt (*SalamonGomez-adaptedparam*). We also observe a voicing false alarm rate of 21%, i.e. segments where the guitar is detected as predominant melody. The raw pitch accuracy for voiced frames is 67.4%.

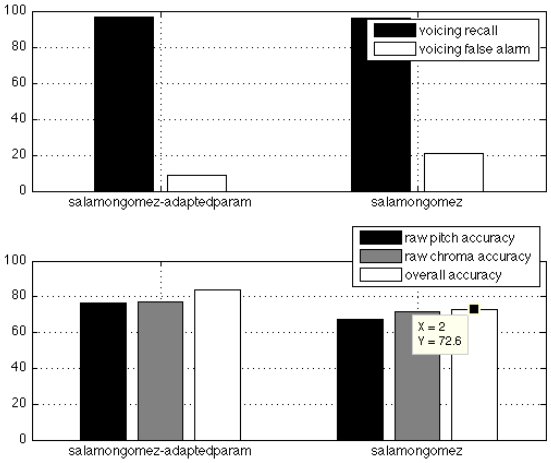

Figure 7: Frame-based accuracy measures (50 cents tolerance) for flamenco singing with guitar accompaniment.

4.3 Error Analysis

We observe that many of the transcription errors appear because of two main factors. The first is the presence of errors in the f0 estimation. In a cappella singing, errors appear when analyzing noisy recordings and reverberation. This is common for all the tested approaches. In addition, we found that the *yin* algorithm had some octave errors in several of the analyzed excerpts. When considering singing voice with guitar accompaniment, some octave errors appear when the guitar is very loud (e.g. at the end of phrases) and for some particular voice timbres. This confirms the importance of good f0 estimation and voicing detection as key steps in automatic transcription systems.

The second factor that influences the final accuracy is the note segmentation method. Fhe fact that tuning is variable over time (a cappella singing) generates wrong note segmentations and labeling. Finally, in many cases the note segmentation algorithm does not correctly segment short notes; either they are consolidated while the annotation consists of several close notes, or vice versa. This especially happens in polyphonic recordings, where the energy envelope also measures the presence of guitar.

4.4 Application Contexts

The melodic transcriptions generated by the proposed method have been used in two different application contexts: melodic similarity and ornament detection. For measuring melodic similarity, the transcriptions are first post-processed. Short notes are consolidated and note pitches were converted into interval values. The obtained melodic contours are compared using standard melodic similarity metrics and evaluated for style classification (Escobar et al. 2008).

The second application context analyzes the obtained transcriptions to detect frequent and representative ornamentation (*melisma*) in flamenco singing. It is based on strategies for pattern detection which consider fundamental frequency and note pitch and duration information (Gómez et al. 2011). Here, the computed transcriptions were only post-processed in order to convert pitch to interval values.

5. Conclusions and Future Perspectives

This paper proposes an approach for computer-assisted transcription of flamenco a cappella singing. We have analyzed the main technological challenges, and proposed an approach based on an iterative note segmentation and labelling technique from f0, energy and timbre. The approach has been evaluated on a collection of annotated performances, obtaining satisfactory results for different f0 estimation algorithms for both monophonic and polyphonic material, comparable to a state-of-the-art approach for note segmentation. We also observed that the main limitations were due to the f0 estimation (e.g. robustness against reverberation and noise of monophonic f0 estimation), voicing detection, tuning, and short note segmentation. Our approach has been further used for comparing performances, styles and variants by means of melodic similarity and for locating frequent ornamentation.

We have seen that there is still much room for improvement. One limination of this work is the small amount of manual annotations, especially for accompanied singing. This is due to the fact that manual annotation is very time consuming and difficult. We are currently expanding the amount of manual annotations. Another limitation is the subjectivity of the task. To address this, we plan to compare annotations by independent experts as a way to quantify the uncertainty of the ground truth information and adapt the algorithm parameters accordingly.

6. Acknowledgements

The authors would like to thank the COFLA team for providing the data set and expert knowledge in flamenco music. We also thank Micheline Lesaffre and authors of (Mulder et al., 2003) for granting access to the *mami* executable. This work has been partially funded by AGAUR (mobility grant), the COFLA project (P09-TIC-4840 *Proyecto de Excelencia, Junta de Andalucía*) and the *Programa de Formación del Profesorado Universitario* of the *Ministerio de Educación de España.*

References

Bregman, A. (1990). Auditory Scene Analysis. MIT Press, Cambridge, Massachussetts, 1990.

Blas Vega, J.; Ríos Ruiz, M. (1988) Diccionario enciclopédico ilustrado del flamenco. Cinterco. Madrid.

Cano, P. (1998). Fundamental Frequency Estimation in the SMS analysis. International Conference on Digital Audio Effects (DAFX).

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, 111, pp. 1917-1930.

Donnier, P (1997). Flamenco: elementos para la transcripción del cante y la guitarra, Proceedings of the IIIrd Congress of the Spanish Ethnomusicology Society.

Escobar, F.J., Díaz-Báñez, J.M., Gómez, E., Mora, J., and Cabrera, J. (2008). Comparative Melodic Analysis of a Cappella

Flanagan, J. L. and Golden, R. M. (1966). Phase Vocoder. Bell Systems Technical Journal, 45:1493–1509, 1966.

Flamenco Cantes, Proceedings of the Conference on Interdisciplinary Musicology, CIM08.

Fernández, L (2004). Flamenco music theory. Acordes concert.

Gómez, E. Klapuri, A. Meudic, B. (2003). Melody Description and Extraction in the Context of Music Content Processing, Journal of New Music Research Vol.32 .1

Gómez, F., Pikrakis, A., Mora, J., Díaz-Báñez, J. M., Gómez, E., Escobar, F. (2011) Automatic detection of ornamentation in flamenco music, 4th International Workshop on Music and Machine Learning: Learning from Musical Structure, Neural Information Processing Systems Foundation, Granada, December 2011.

Hurtado Torres, D. and Hurtado Torres A. (1998) El arte de la escritura musical flamenca. Bienal de Arte Flamenco. Sevilla.

Hurtado Torres, A. and Hurtado Torres, D. (2002). La voz de la tierra, estudio y transcripción de los cantes campesinos en las provincias de Jaén y Córdoba. Centro Andaluz de Flamenco.

Jerez.
Hoces, R. (2011). La transcripción musical para guitarra flamenca: análisis e implementación metodológica. PhD thesis, PhD program "Estudios avanzados de flamenco: un análisis multidisciplinar", University of Seville.
Israel J. Katz. Flamenco. Grove Music Online ed. L. Macy (Accessed 16 May, 2006), www.grovemusic.com
Janer, J., Bonada, J., de Boer, M., Loscos, A. (2008). Audio Recording Analysis and Rating, Patent pending US20080026977, Universitat Pompeu Fabra, 06/02/2008.
Klapuri, A. and Davy, M. (Editors) (2006). Signal Processing Methods for Music Transcription. Springer-Verlag, New York.
Lesaffre, M., Leman, M., De Baets, B. and Martens, J.-P. (2004). Methodological considerations concerning manual annotation of musical audio in function of algorithm development, Proceedings of the International Conference on Music Information Retrieval.
Maher, R. C. and Beauchamp, J. W. (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure, Journal of the Acoustical Society of America, Vol. 95(4), pp. 2254-2263.
MIREX wiki, http://www.music-ir.org/mirex/wiki/, accessed November, 14, 2011.
Mora, J., Gomez, F., Gómez, E., Escobar-Borrego, F.J., Diaz-Bañez, J.M. (2010). Characterization and melodic similarity of a Cappella flamenco cantes. Proceedings of the International Society for Music Information Retrieval Conference.
Mulder, T., Martens, J. P. Lesaffre, M., Leman, M., De Baets, B., De Meyer, H. (2003), An Auditory Model Based Transcriber of Vocal Queries, Proceedings of the International Conference on Music Information Retrieval.
Navarro, J.L., Ropero, M. (editor) (1995). Historia del flamenco. Ed. Tartessos, Sevilla.
Ryynänen, M. P. (2006). Singing transcription, in Signal processing methods for music transcription (A. Klapuri and M. Davy, eds.), Springer.
Salamon, J.,Gómez, E. and Bonada, J. (2011). Sinusoid Extraction and Salience Function Design for Predominant Melody Estimation. In Proc. of the 14th International Conference on Digital Audio Effects (DAF-x 11), Paris, France, September 2011, pp. 73-80.
Salamon, J. and Gómez, E. (2011). Melody Extraction from Polyphonic Music: MIREX 2011. In 5th Music Information Retrieval Evaluation exchange (MIREX), extended abstract, Miami, USA, October 2011.
Salamon, J. and Gómez, E. (2012). Melody Extraction from Polyphonic Music Signals using Pitch Contours Characteristics. IEEE Transactions on Audio, Speech and Language Processing, In Press.
Sundberg, J. (1987). The Science of the Singing Voice. DeKalb, IL: Northern Illinois Univ. Press.
Toiviainen, P. & Eerola, T. (2006). Visualization in comparative music research. In A. Rizzi & M Vichi (Eds.), COMPSTAT 2006 - Computational Statistics. Heidelberg: Physica-Verlag, 209-221.
Vickers, E. (2001). Automatic Long-Term Loudness and Dynamics Matching. In Proc. of the Conv. of the Audio Engineering Society (AES), 2001.